Final Technical Report

Award Number: 03HQGR0004

Recipient: Columbia University, NY

Principal Investigator: Felix Waldhauser

Title:

Improved Differential Travel Time Measurements and a Search for Repeating Events at the Northern California Seismic Network

Program Element: I
Keywords: Seismology, Source characteristics, Database

February 24, 2004

Research supported by the U.S. Geological Survey (USGS), Department of the Interior, under USGS award number 03HQGR0004. The views and conclusions contained in this document are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of the U.S. Government.

Award Number: 03HQGR0004

Improved Differential Travel Time Measurements and a Search for Repeating Events at the Northern California Seismic Network

Felix Waldhauser
Lamont-Doherty Earth Observatory of Columbia University
Palisades, NY 10964
Tel: (845) 365 8538; Fax: (845) 365 8150;
felixw@ldeo.columbia.edu

TECHNICAL ABSTRACT

We processed the complete digital seismogram data base for Northern California to measure accurate differential times for correlated earthquakes observed at common stations. Correlated earthquakes are earthquakes that occur within a few kilometers of one another and have similar focal mechanisms, thus generating similar waveforms, which allows measurements to be made via cross correlation analysis. The waveform data base was obtained from the Northern California Earthquake Data Center and includes about 15 million seismograms from 225,000 local earthquakes between 1984-2001. A total of 26 billion cross correlation measurements were performed on a 32-node (64 processor) Linux cluster, using improved analysis tools. We computed a total of about 1.7 billion P-wave differential times from pairs of waveforms with cross correlation coefficients of 0.6 or larger. The P-wave differential times are on the order of a factor of ten to a hundred times more accurate than those obtained from routinely picked phase onsets. 1.2 billion S-wave differential times were measured, a phase not routinely picked at the Northern California Seismic Network because of the generally weak onset of S-phases. We found that approximately 90% of the seismicity includes events that have cross-correlation coefficients of CC>0.7, with at least one other event recorded at four or more stations. At some stations more than 40% of the recorded events are similar at the CC>0.9 level, indicating the potential existence of large numbers of repeating earthquakes. Large numbers of correlated events occur in different tectonic regions, including the San Andreas Fault, the Long Valley Caldera, and the Mendocino Triple Junction. Future research using these data may substantially improve earthquake locations and add insight into the velocity structure in the crust. Such work will help the seismic hazard community to better understand the risks and hazards associated with earthquakes in Northern California.

NON-TECHNICAL ABSTRACT

We processed the complete seismogram data base for Northern California to measure accurate differential travel times for correlated earthquakes observed at common stations. Correlated earthquakes are earthquakes that occur within a few kilometers of one another and have similar focal mechanisms, thus generating similar waveforms, which allows measurements to be made via cross correlation analysis. The new data improves by a factor of ten to hundred times the accuracy of a fundamental measurement in seismology. Inspection of the new data reveals that a high percentage of earthquakes in Northern California are similar, with many of them likely to be of repeating character. Future research using these data may substantially improve earthquake locations and our knowledge of the velocity structure in the crust, which eventually will help to reduce earthquake hazard in Northern California.

Introduction and overview

One of the most fundamental datasets in seismology is the set of measured arrival times of various phases on a seismogram. These basic data are used to solve for earthquake hypocenters and also to derive velocity models or travel time curves. But there is an error associated with each measurement. Average pick errors for Northern California Seismic Network (NCSN) phase data are on the order of 0.1 sec. These errors map into significant scatter in the earthquake locations and reduce the resolution of tomographic inversions.

It has long been established that cross correlation measurements of differential travel times can improve these errors by an order of magnitude or more if the waveforms are similar. Such similar waveforms are produced when earthquakes have the same rupture mechanism and are co-located (i.e. share the same ray paths between source and receiver). Figure 1 shows an example of 38 virtually identical waveforms of a repeating earthquake source recorded at the NCSN station JST. For such similar waveforms, relative phase arrival times can be obtained with sub-sample precision (Poupinet et al., 1984). With a sampling rate of 100 samples/second, errors in relative arrival time measurements are less than 1 ms in the optimal case. In comparison, absolute phase arrival times are picked at the NCEDC with an accuracy of about 0.1 s on average. Even when earthquakes are not exactly co-located, waveforms can be similar enough to provide significant improvement in relative arrival time measurements. Cross correlation measurements are particularly important for S-waves, because the onsets are often obscured by the P-wave coda and as a result are rarely picked for NCSN data. The substantial number of added S-wave measurements will provide better constraint on earthquake locations and are especially important for resolving the depth.

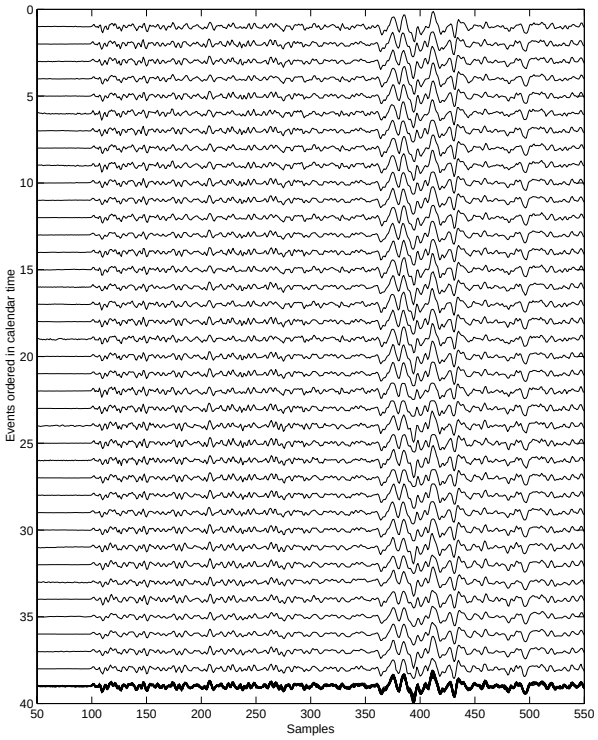

Figure 1 Waveforms of 38 repeating earthquakes on the Calaveras fault recorded at station JST. Bottom trace shows all 38 waveform superposed. Note the high similarity between all waveforms. From Schaff (2001).

Here we present the final report of a one-year project, describing the work undertaken by the principle investigator Felix Waldhauser and, for the larger part, by David Schaff (post doctoral research scientist). Our research included the collection and reformatting of all digital waveform data recorded by the Northern California Seismic Network (NCSN), the development of computational tools to measure differential travel times by waveform cross correlation on a massive scale, the application of these tools to the entire NCSN waveform data base, and an initial investigation of the new data. Future work using these data is expected to produce substantially more precise event locations across large areas, which has the potential to lead to significant new insights into the structure, mechanics, and dynamics of fault systems, and to a better assessment of earthquake hazard.

Investigations undertaken

We computed highly accurate differential travel times from waveform cross correlation for all similar events observed at common stations between 1984-2001 across the Northern California Seismic Network (NCSN). We performed about 26 billion correlation measurements on 225,000 local events, resulting in a total of 1.7 billion P-wave and 1.2 billion S-wave differential times. We performed an initial analysis of the new data to investigate its characteristics and its potential use in future research.

The following milestones were achieved to complete this enormous computational task:

Collection, local storage, and reformatting of the entire NCSN waveform data base

Doug Neuhauser at the Northern California Earthquake Data Center (NCEDC) kindly extracted all of the 225 GB of waveform data to 10 DLT tapes (~700 GB in uncompressed format). The data is stored in a compressed binary CUSP format which we converted to SAC file format for processing. The seismograms were reorganized from a calendar ordering to a station ordering, as correlation measurements are performed on a station by station basis. To accomplish this, it was most expedient to have enough disk space to accommodate all 225 GB at one time. For this we purchased two internal 120 GB hard drives from the funds of this project. The seismograms were uncompressed, preprocessed with travel time information for P- and S-waves being updated to the event headers, and then recompressed. These operations were performed for 15 million SAC files or seismograms. Data transfer rates from disk and across our network are on the order of 1 MB/sec. This amounted to about 3 days of computer time for each disk access operation if uninterrupted — copying the data from the DLT drive, uncompressing, converting from CUSP to SAC, recompressing, and reorganizing into station subdirectories. The DLT drive, however, can only be manually operated and the tapes must be changed after each extraction is complete. The total amount of time involved for data handling and manipulation, development of software, and testing the integrity of the transfer and conversion amounted to about 2 months.

Development of computational tools for massive scale cross-correlation

Our earlier work used cross correlation routines that were designed to satisfy the memory and speed requirements for processing on the order of 10K events (Schaff et al., 2002; Schaff et al., 2003). With the task of processing over an order of magnitude more data, these routines needed to be modified to efficiently process larger numbers of events recorded by a single station. The initial correlation program used FORTRAN subroutines for the number crunching and MATLAB to facilitate the bookkeeping. To improve both the memory and speed, we have converted the whole

program to FORTRAN and added some new features. Resampling and filtering is now performed on-the-fly internally within core memory. Also, a toggle feature for byte swapping has been included so that data can be analyzed on both Sun and Linux platforms. Depending on the window lengths and the lags searched over, we are able to perform about 10 million correlation measurements per hour, a factor of 10 improvement in speed over our earlier routines.

We have experimented with a correlation detector which is able to recover lags greater than half the window length. This is a new feature and different than the correlation function which was applied in our earlier work. Figure 2 shows examples of automatically determined P-wave arrival time adjustments of similar events observed at station JST. These P-wave trains have CC > 0.9 and adjustments > 0.9 sec for window lengths of 1 sec. All of these event pairs had at least one theoretical initial window alignment, which is the reason for the large adjustments. They would have been missed with more standard methods of cross correlation.

Computations were performed on a 32 node Linux cluster, each node equipped with two 1.2 GHz Athlon processors, 1 GB of fast RAM, and 20 GB of scratch space. A RAID Tb storage system was used to store the waveforms and the measurement output. Since the correlations operate on a station by station basis, they are naturally parallelizeable and can use any number of free processors. Cross correlations are performed at a rate of about 10 million measurements per CPU hour.

Event pair selection based on double-difference locations

Since waveform similarity breaks down with increasing inter-event separation distance, we implemented an event separation threshold to select event pairs suitable for cross correlation. To improve the accuracy of inter-event distances from which we determine such pairs of events we have relocated about 240,000 events using the double-difference algorithm hypoDD (Waldhauser and Ellsworth, 2000; Waldhauser, 2001) together with about 5 million NCSN P-phase picks. Using these improved locations, we chose an inter-event distance threshold of 5 km. The somewhat large separation distance is chosen to account for remaining errors in the relative locations, in particular between events not relocated by the double-difference method due to lack of good station coverage, and to find as many similar events as possible. In addition, from a quarter wavelength rule we don't expect events separated by greater distances to correlate well (Geller and Mueller, 1980).

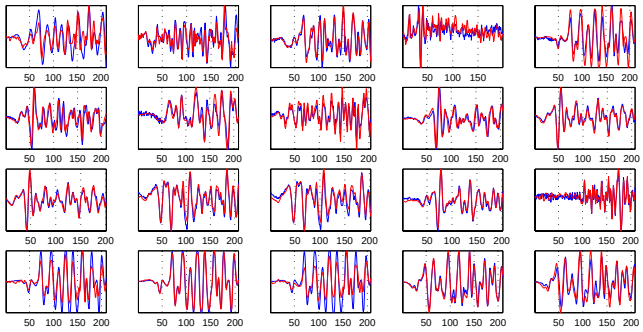

Figure 2 A few examples of aligned P-waves for several pairs of events (blue and red overlaying seismograms) obtained from a correlation detector. All adjustments are > 0.9 sec which is more than half the window length of 1 sec. The P-wave trains are very similar with CC > 0.9. X-axes are in samples (delta = 0.01 sec.).

Initial test runs and performance evaluation

Before embarking on massive processing of all the NCSN waveform data, we executed a battery of tests to evaluate the performance of the correlation method and judge which parameters give the best results uniformly across diverse tectonic settings. We also examined the statistics of various parameters involved with the correlation to help gauge appropriate thresholds, values, and judge quality in order to remove outliers before the location. We explored the use of filtering (filter bands of 1.5 to 15 Hz) and resampling to obtain more usable correlation measurements, as well as the incorporation of theoretical P- and S-wave initial window alignments, if no phase picks were available. We experimented with different event separation distance cut-offs and using improved double-difference locations from phase data to determine the initial pairs. We also examined the effects of different window lengths (1 and 2 sec) on the correlation coefficients (CC) and the robustness of the delay measurements.

Results

Cross correlation and delay time measurements

About 26 billion P- and S-wave differential times with their respective cross correlation coefficients were calculated during the course of this project. This actually represents 13 billion correlation measurement pairs since both windows of 1 and 2 sec were computed. We saved all data with correlation coefficients (CC) of 0.6 and above, amounting to 63 GB of output files, with 1.7 billion P-wave and 1.2 billion S-wave correlation pair measurements for both window lengths.

Figure 3 shows all events recorded at the NCSN between 1984 and 2001 that have similar P-wave trains at the CC>0.7 level with at least one other event at four or more stations. The ~200,000 events represent 90% of the total number of events for which waveforms are available. 80% of the total number of events share similar waveforms with at least one other events at eight or more stations, and 75% at 10 or more stations. These surprisingly high numbers indicate that a large percentage of the NCSN catalog can be relocated with differential travel times obtained from waveform cross correlation.

Areas with large numbers of highly correlated events can be identified in Figure 4. This figure shows the percentage of events, within bins of 5x5 km, that have CC>0.7 with at least one other event at four or more stations. It indicates that greater than a 75% level is obtained for much of the area in different tectonic settings such as the Long Valley Caldera, Bay area faults, the Geysers Geothermal Field, and the Mendocino Triple Junction. The large percentages in these areas are related to the dense distribution of seismicity that produce similar waveforms for many closely located events.

A similar picture is obtained when the percentage of correlated events is plotted for individual stations that recorded them (Figure 5). The stations that recorded many correlated events are located along the SAF, at Geysers Geothermal Field, and in the Long Valley area (Figure 5a). Figure 5b indicates the percentage of cross correlation measurements that have P-wave correlation coefficients of CC>0.7. Again, stations that recorded events from creeping faults along the SAF system (e.g., Calaveras or the Parkfield section of the SAF) have significantly larger percentages of correlated waveforms than stations that record seismicity in areas that are seismically less active.

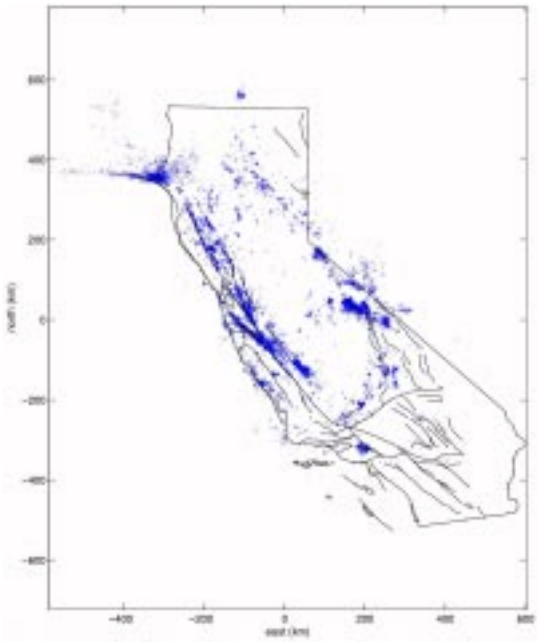

Figure 3 All events between 1984 and 2001 that have P-wave cross correlation coefficients of CC>0.7 with at least one other event at four or more stations. These 200,000 events represent 90% of the total seismicity with waveforms available.

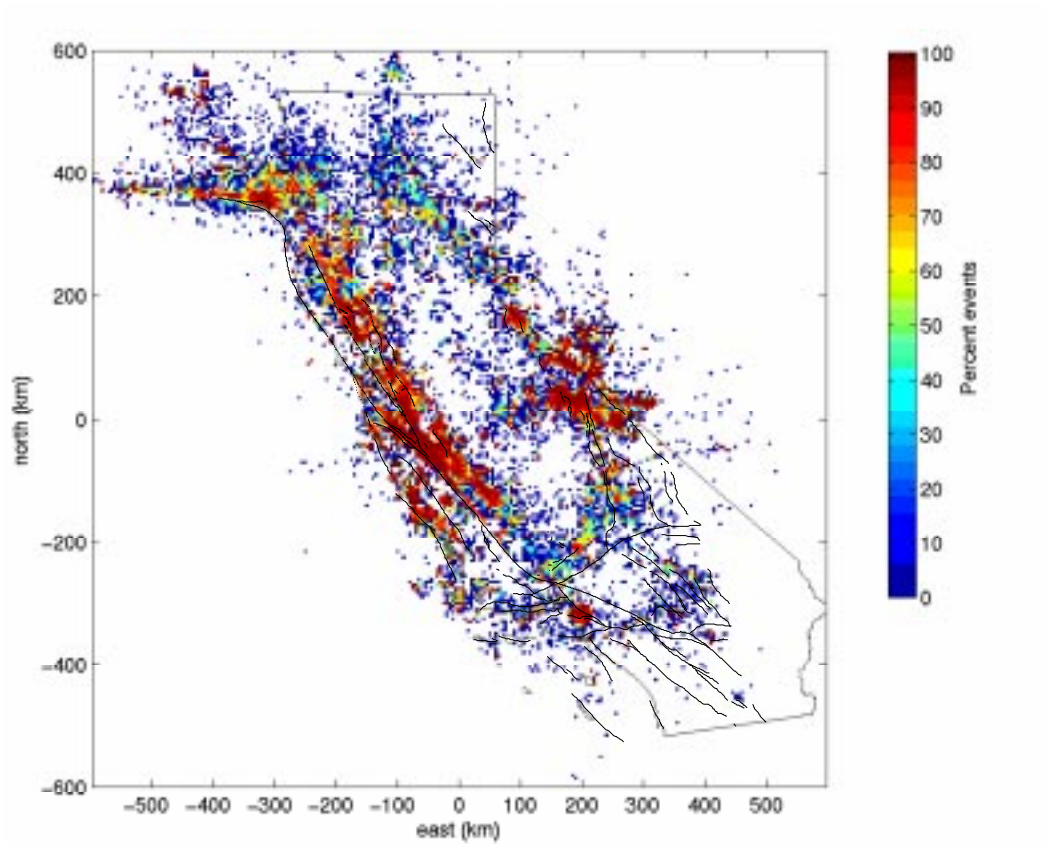

Figure 4 Percentage of correlated events that have cross-correlation coefficients of CC>0.7 with at least one other event recorded at four or more stations. Percentages are computed from the total number of events within bins of 5x5 km.

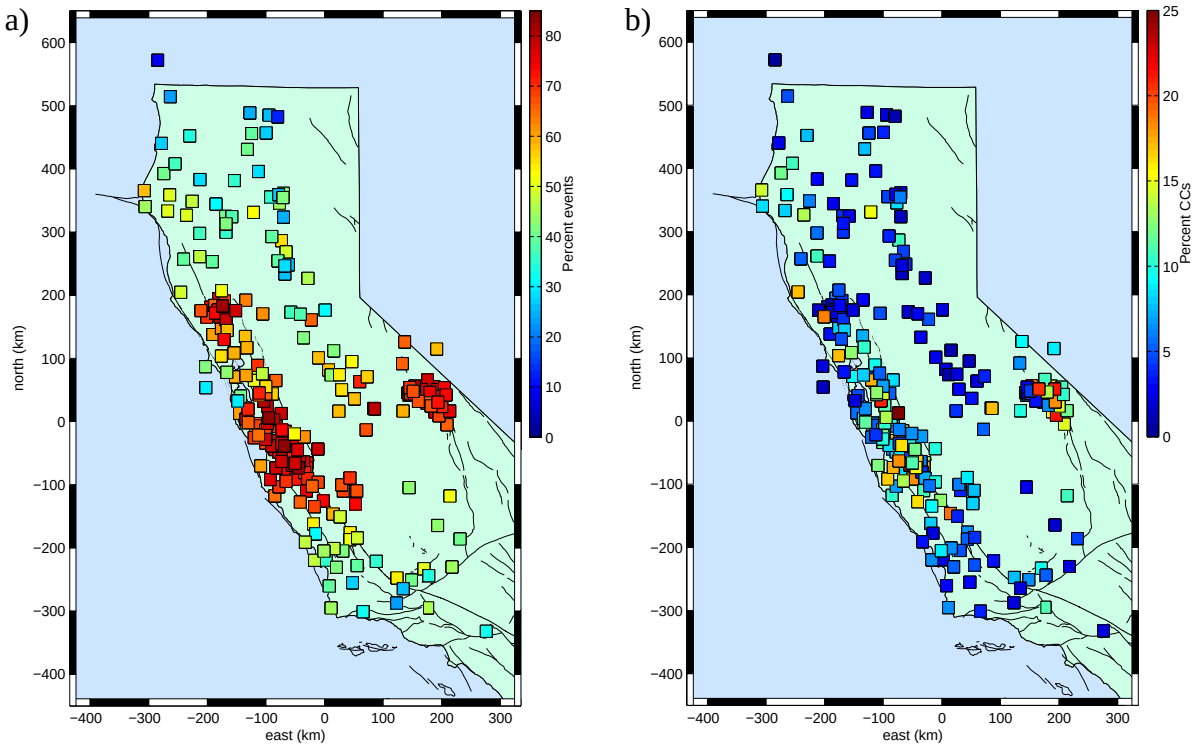

Figure 5 Station locations with percentages indicated for a) number of correlated events recorded at each station, and for b) number of cross correlation measurements with CC>0.7 performed.

Three example stations: JSF, MDR, and KBB

Figure 6 shows the detailed results for station JST. At this particular station, which recorded 35,000 events from the San Andreas Fault system, 40% of the events have at least one other event with cross correlation coefficients (CC) greater than 0.9 (62% for CC > 0.8, 77% for CC > 0.7) (Figure 6b). The percentages of similar events observed at station JST are surprisingly high, but they include known areas of repeating events on the Calaveras and San Andreas Faults. For a station in Long Valley Caldera (MDR) recording 72,000 events, the distribution is 18% for CC > 0.9, 43% for CC > 0.8, and 67% for CC > 0.7. A station including 20,000 events in the different tectonic settings of Mendocino Triple Junction and Geysers Geothermal Fields yields correlation measurements where 16% of the events have at least one other event with CC > 0.9, 32% with CC > 0.8, and 49% with CC > 0.7. The lower numbers of correlated events observed at the latter two stations most likely reflect the different faulting processes that take place in these areas, compared to the (mostly) strike-slip events recorded at JST. Table 1 and Table 2 summarize the number of measurements for the three stations and different thresholds. It is seen from Table 2 that a large percentage of the events correlate well across a variety of tectonic regions.

There are virtually no S-wave picks in the phase data for the 240,000 events at the NCEDC. By using theoretical initial window alignments based on 1.732 times the P-wave travel time and performing cross correlations on windows containing S-wave energy, we are able to obtain nearly the same number of S-wave observations as for P-waves (Figure 6b, Tables 1 & 2). We are there-

fore able to nearly double the number of observations that can be used for future location by including S-waves that haven't been picked and also benefit from the extra constraint they provide on depth. Filtering is also seen to increase the number of useful measurements (Figure 6b). Not all the seismograms associated with an event have P-wave picks perhaps due to weak onsets or low signal-to-noise ratios. If we use theoretical initial P-wave window alignments based on raytracing through a 1D velocity model, we are also able to increase the number of observations by about 30% compared to if we only used event pairs that had P-picks for both events listed in the NCSN bulletin (see Table 1).

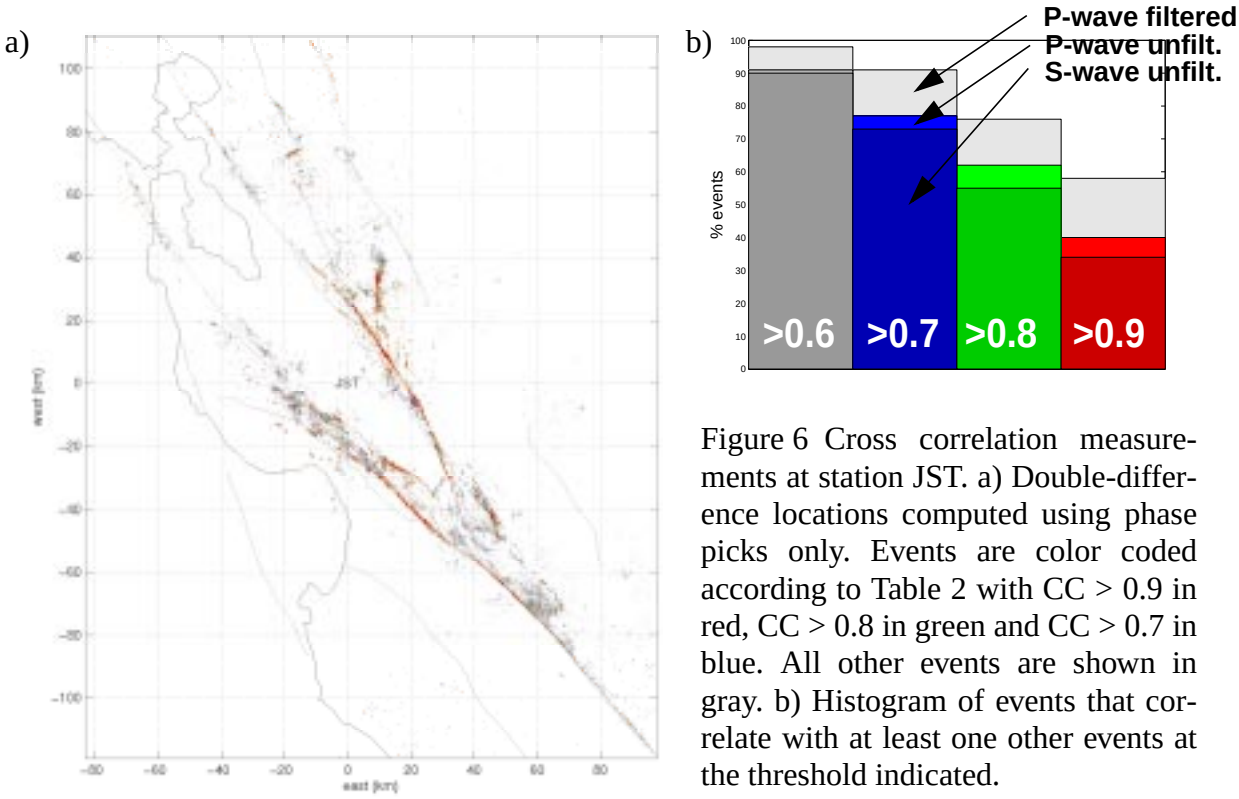

Figure 6 Cross correlation measurements at station JST. a) Double-difference locations computed using phase picks only. Events are color coded according to Table 2 with CC > 0.9 in red, CC > 0.8 in green and CC > 0.7 in blue. All other events are shown in gray. b) Histogram of events that correlate with at least one other events at the threshold indicated.

Table 1: Number of Correlation Measurements

	CC > thresh			
station (phase)	0.6	0.7	0.8	0.9
JST (P-wave)	1.3 M (7%)	495 K (3%)	165 K (0.9%)	43 K (0.2%)
MDR (P-wave)	5.1 M (5%)	1.5 M (1%)	355 K (0.3%)	29 K (0.03%)
KBB (P-wave)	293 K (21%)	114 K (8%)	38 K (3%)	9 K (0.7%)
JST (S-wave)	1.7 M (9%)	656 K (3%)	215 K (1%)	54 K (0.3%)
JST (theor P-wave)	308 K (30%)	105 K (28%)	36 K (27%)	10 K (31%)
JST (P-wave filtered)	4.1 M (21%)	1.7 M (9%)	578 K (3%)	136 K (0.7%)

Table 2: Number of Events

	CC > thresh			
station (phase)	0.6	0.7	0.8	0.9
JST (P-wave)	32 K (91%)	27 K (77%)	22 K (62%)	14 K (40%)
MDR (P-wave)	58 K (81%)	483 K (67%)	31 K (43%)	13 K (18%)
KBB (P-wave)	14 K (78%)	10 K (57%)	6 K (36%)	3 K (19%)
JST (S-wave)	31 K (90%)	25 K (73%)	19 K (55%)	12 K (34%)
JST (P-wave filtered)	34 K (98%)	32 K (91%)	27 K (76%)	20 K (58%)

Characteristics of the cross correlation data

Figure 7a shows the contours of the distribution of CC vs. inter-event separation distance for station JST. It decreases as expected because of the breakdown in waveform similarity with increasing separation. The different confidence levels are shown in the legend. They are computed by dividing the x-axis into 1000 bins of equal number represented by each point (e.g. JST has 1900 obs per bin). From figures like this we were able to determine that event separations of 5 km and less should probably capture most of the useful cross correlation measurements.

Using correlation coefficient thresholds is currently the primary means for deciding what data to include for future location studies. We sought additional independent means to judge measurement quality and remove outliers. Computing correlations at two different window lengths provides two independent relative arrival time measurements that should agree for the same phase at the same station. Figure 7b shows the distribution of the difference in absolute adjustments for two window lengths, abs(dt2-dt1). For station JST, which has lots of similar events, the values agree to two samples (0.02 sec) or better all the way out to CC = 0.6. Combined with CC thresholds this can be an additional way to remove measurement outliers. From such a procedure we were also able to determine that filtering can remove some large outliers associated with long period instrument noise even though the correlation coefficients were high and therefore not excluded on that basis.

Another interesting aspect we explored with the new data is the degree with which crustal heterogeneity between source region and recording stations controls waveform similarity. We use a subset of about 1500 precisely located events along the Parkfield section of the San Andreas Fault. For each event pair/station configuration for which a cross correlation coefficient of 0.7 or larger is obtained, we determine inter-event distance and the azimuth between the direction of the event pair and the station that recorded both events. Figure 8 shows the variation of these cross correlation coefficients as a function of the event pair/station azimuth and different intervals of event separation. As expected, for events that are co-located, the cross correlation coefficients are insensitive to variation in recording azimuth. With increasing recording azimuth, and within intervals of inter-event distances, we observe a trend of CC decrease. This is because at zero azimuth the rays for two events travel a similar path outside the source region since the station is in direction of the event pair. At an azimuth of 90° ray paths are perpendicular to the direction of the event pair, and thus travel through increasingly different media, compared to the case where the event separation and slowness vectors are parallel (0°).

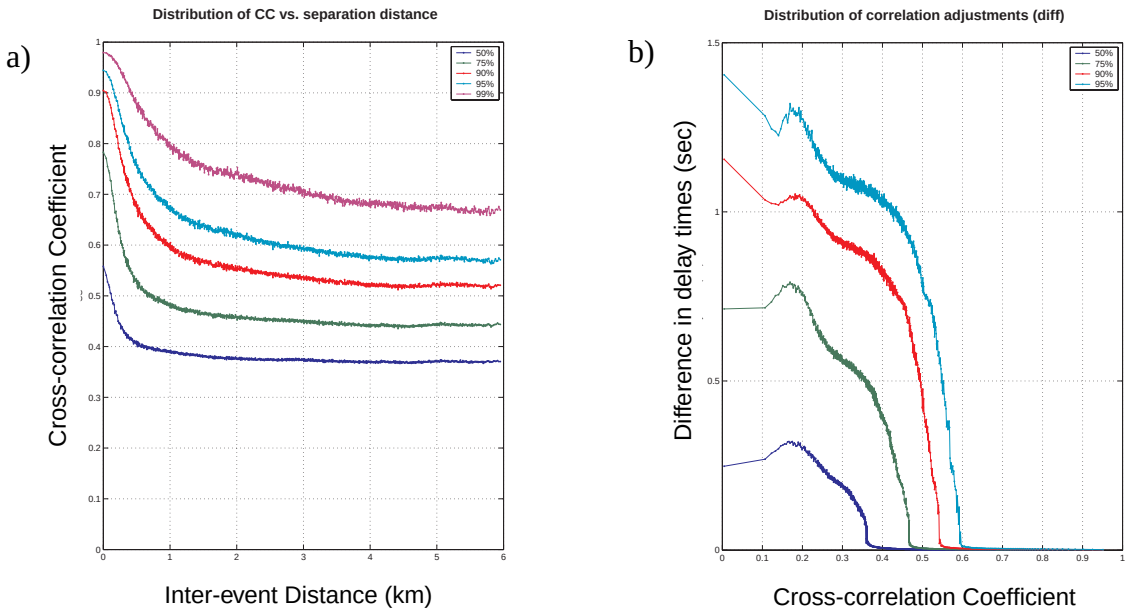

Figure 7 Statistics of correlation measurements for station JST. (a) Breakdown of similarity with event separation distance. (b) Agreement of delay measurements for two different window lengths.

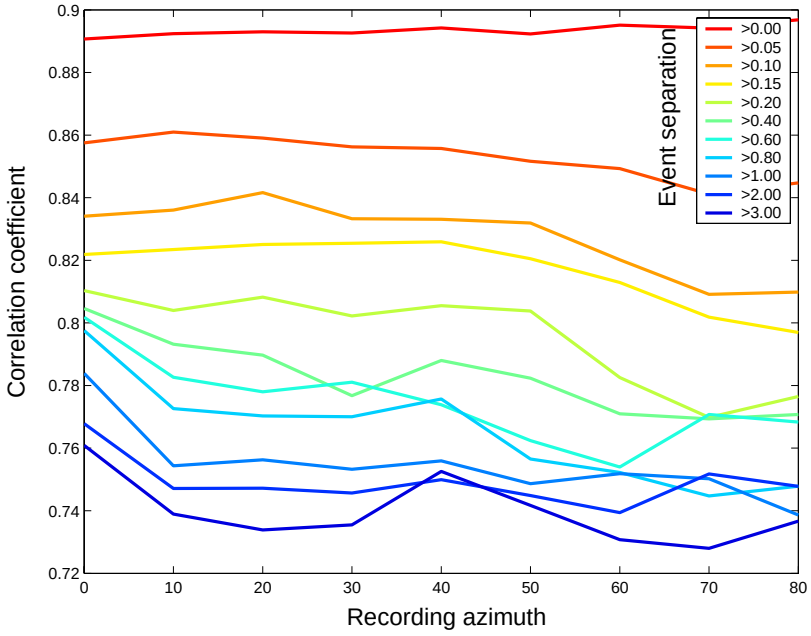

Figure 8 Cross-correlation coefficients as a function of recording azimuth (station relative to direction of event pair), for different intervals of inter-event distances (km). Data are from 1500 events on the San Andreas Fault.

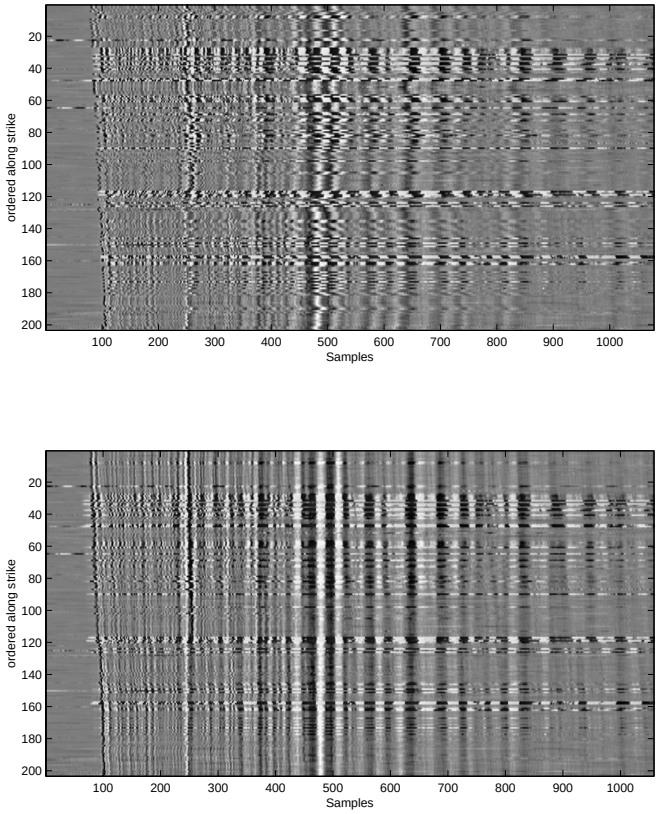

Figure 9 Unfiltered waveforms of earthquakes along a 2 km long streak on the Calaveras fault recorded at station CCO. Seismograms are ordered by distance and aligned (top) according to the catalog P-wave picks and (bottom) by cross correlation. The scatter in the phase picks is removed by more than an order of magnitude.

The improvement that correlation measurements can make compared to catalog P-wave picks is shown in Figure 9 for 200 events along a 2 km long streak on the Calaveras Fault, recorded at one station. The 200 waveforms are highly similar and are shown before (top panel) and after alignment (bottom panel). The moveout of the P-wave relative to the S-wave in the bottom panel is appropriate for a 2 km offset. Alternating black and white bands for events 25-42 in the top panel are seen to be exactly out of phase or misaligned by about 10 samples. The same waveforms in the bottom panel are aligned to the nearest sample or better and therefore represent at least an order of magnitude improvement in the differential travel times. Note that no S-phases are routinely picked at the NCSN, while differential S-wave times can be measured at sub-sample precision.

Concluding remarks

This project resulted in a wealth of cross correlation and differential time information, consistently measured for all events digitally recorded by the Northern California Seismic Network between 1984 and 2001. The data is uniformly computed across Northern California, allowing

direct comparison between different tectonic areas including NEHRP priority faults such as the Hayward, Calaveras, and the San Andreas fault, as well as other hazard areas like Long Valley Caldera. Preliminary inspection of these data indicates its potential usefulness in a wide range of future research, including but not limited to regional scale earthquake relocation studies and tomographic investigations, as well as characterization of crustal heterogeneity across Northern California. Double-difference relocations of the NCSN catalog with only the phase data showed a substantially increased level of detail across most of the Northern California region (see e.g. Figure 6), which can be significantly enhanced by incorporating the cross correlation differential times measured during this project.

We note that this project had significant slow downs due to delays in delivery and installation of the RAID file server which was essential to process the waveforms efficiently. We were nevertheless able to complete the measurements in time. We have initially proposed to carry out about 2.5 billion cross correlation measurements, but by the end of this project we carried out about 26 billion measurements in order to create a data base that is as comprehensive as possible. We were positively surprised by the tremendous amount of correlated events that exists in the catalog of the NCSN. An initial search for repeating events based on these new data, as stated in the proposal, is under way but prolonged due to the unexpected size of the useful data output. We will also work towards a data base which can be easily accessed by interested researchers.

The NCSN waveform data used in this project is freely available at the NCEDC at UC Berkeley. Since a fair amount of effort is required to change the data from an event (calendar time) ordering scheme to a station based ordering, it is possible that we could make the reorganized 225 GB dataset available to interested researchers or even to the data center itself. Due to funding level constraints, it was agreed in the revised budget for this project to make the database openly and publicly available within a year after completion of the project, because of the large potential benefit to the greater geophysical community.

References

Geller, R. J. and C. S. Mueller (1980). Four similar earthquakes in Central California, *Geophys. Res. Lett.* 7, 821-824.

Poupinet, G., W.L. Ellsworth, and J. Fréchet (1984). Monitoring velocity variations in the crust using earthquake doublets: an application to the Calaveras fault, California, *J. Geophys. Res.* **89**, 5719-5731.

Schaff, D. P., G. H. R. Bokelmann, W. L. Ellsworth, E. Zanzerkia, F. Waldhauser, and G. C. Beroza. Optimizing correlation techniques for improved earthquake location, *Bull. Seism. Soc. Am.*, in press.

Schaff, D.P., 4D high resolution seismology: repeating events and large scale relocation, *Ph.D. Thesis*, Stanford University, 115pp., 2001.

Waldhauser, F., and W.L. Ellsworth, A double-difference earthquake location algorithm: method and application to the northern Hayward Fault, California, *Bull. Seism. Soc. Am., 90*, 1,353-1,368, 2000.

Waldhauser, F., HypoDD: A computer program to compute double-difference hypocenter locations, *U.S.G.S. open-file report*, 01-113, Menlo Park, California, 2001.

Reports published (related to this project)

Schaff, D. and F. Waldhauser, Waveform cross correlation and differential time measurement at the Northern California Seismic Network, *in preparation for BSSA, 2004.*

Schaff, D., F. Waldhauser, and P.G. Richards, Applying Massive Waveform Cross Correlation and Double-Difference Location to Northern California and China, *Eos Trans. AGU, 84*(46), Fall Meet. Suppl., 2003.

Schaff, D. and F. Waldhauser, Progress in massive waveform cross correlation and wide area event relocation in Northern California, *Proceedings and Abstracts*, Volume XIII, Southern California Earthquake Center Annual Meeting, Oxnard, CA, Sept. 7-11, 2003.

Waldhauser, F. and D. Schaff, Cross-correlation and double-difference relocation in Northern California, presented at the SCEC Workshop on Converting Advances in Seismology into Earthquake Science, Caltech, Pasadena, September 22-23, 2003.